Small-scale inference: Empirical Bayes and confidence methods for as few as a single comparison

January 18, 2013

David R. Bickel

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology, and Immunology

University of Ottawa; 451 Smyth Road; Ottawa, Ontario, K1H 8M5

Abstract

By restricting the possible values of the proportion of null hypotheses that are true, the local false discovery rate (LFDR) can be estimated using as few as one comparison. The proportion of proteins with equivalent abundance was estimated to be about 20% for patient group I and about 90% for group II. The simultaneously-estimated LFDRs give approximately the same inferences as individual-protein confidence levels for group I but are much closer to individual-protein LFDR estimates for group II. Simulations confirm that confidence-based inference or LFDR-based inference performs markedly better for low or high proportions of true null hypotheses, respectively.

Keywords: confidence distribution; empirical Bayes; Lindley's paradox; local false discovery rate; multiple comparison procedure; multiple testing; observed confidence level; restricted parameter space

1 Introduction

In the development of statistical methods for interpreting high-dimensional genomics data, the challenges involved in analyzing genomics data sets of much smaller scale have been largely overlooked, and yet such data are routinely generated. Out of the thousands of genes in the human genome, the expression levels of only on the order of 30 genes are measured in a real-time polymerase chain reaction experiment. Among the hundreds of thousands of proteins in the human proteome, the abundance levels of only on the order of 200 proteins are measured with mass spectrometry. The following idealization of the candidate-gene approach to genetic association studies poses a problem encountered in analyzing data from a small fraction of a large number of biological features, with each feature corresponding to a different population in the sampling theory sense.

Example 1. Consider 10^6 populations such that $X_i \sim \mathrm{N}(\mu_i, 1)$ for $i = 1, \ldots, 10^6$, where $\mu_i = 2$ for N_1 values of i and $\mu_i = 0$ for $10^6 - N_1$ values of i. None of the random values is observed except x_1, the realization of X_1. The null hypothesis of interest is $\mu_1 = 0$. Let Φ and ϕ respectively denote the standard normal distribution function and density function.

Without any knowledge of N_1, few would question the applicability of the p-value $1-\Phi(x_1)$. On the other hand, in the absence of other information, the use of $P(\mu_1 = 0; N_1) = 1 - N_1/10^6$ as an approximate, nonsubjective prior probability of the null hypothesis in order to obtain the approximate posterior probability

$$P(\mu_1 = 0|x_1; N_1) = \frac{(1 - N_1/10^6)\,\phi(x_1)}{(1 - N_1/10^6)\,\phi(x_1) + (N_1/10^6)\,\phi(x_1 - 2)} \tag{1}$$

would not be controversial if N_1 were known. Suppose that N_1 is unknown but can be safely assumed to be between 1 and 100. Then, for at least 99.99% of the populations, the null hypothesis is true and thus $1 - \Phi(X_1) \sim \mathrm{U}(0, 1)$. By contrast, for those same populations, $P\left(\mu_1 = 0|X_1; \tilde{N}_1\right) \approx 1$ with high probability regardless of the value $\tilde{N}_1$ between 1 and 100 that is guessed for N_1 in computing the posterior probability. For instance, if $x_1 = 2$, then the p-value is $1-\Phi(2) = 2.28\%$ even though the posterior probability of the null hypothesis is at least $P(\mu_1 = 0|2; 100) = 99.93\%$ and possibly as high as $P(\mu_1 = 0|2; 1) = 1 - 7.39 \times 10^{-6}$. Lindley (1957) thoroughly examined a similar “paradox” from a more Bayesian viewpoint.

The type of problem faced in Example 1 will be attacked by adapting methodology recently developed for gene expression microarray data to two other settings: (1) those with data available for testing only a much smaller number of hypotheses and (2) those with much smaller proportions of null hypotheses that are true.

Microarray technology enables the measurement of levels of gene expression for thousands of genes in cells under two different conditions, conveniently labeled as treatment and control. Which genes have differential expression in the mean between the treatment and control populations? That large-scale problem of multiple comparisons led Efron et al. (2001) to apply the false discovery rate (FDR) of Benjamini and Hochberg (1995) and to introduce the local false discovery rate (LFDR). In accordance with its name, the LFDR is a rate of Type I errors that would be incurred were the null hypothesis rejected every time the same data are generated as those actually observed. In the microarray context, the LFDR is an empirical Bayes posterior probability of the null hypothesis that a particular gene does not have differential expression, as in equation (1). More precisely, the LFDR is defined as the prior probability of the null hypothesis conditional on the p-value or other statistic that reduces the measured expression levels of the gene to a single number (Efron, 2010b).

Here, like in Example 1, the prior probability approximates an unknown proportion of null hypotheses that are true, with each null hypothesis corresponding to a different gene. In that sense, the LFDR differs from a fully Bayesian posterior probability, which requires the complete specification of the prior distribution of all unknown parameters. Such specification usually involves prior probabilities that correspond to hypothetical levels of belief rather than real relative frequencies or proportions. Thus, whereas a purely Bayesian prior is necessarily known in principle, empirical Bayes priors are unknown.

Since the LFDR generally depends on parameters that do not have a known prior distribution, the LFDR can only be estimated. Supposing, however, that the LFDR could be known and neglecting any information lost in reducing the data to a test statistic for each hypothesis, Bayes decision rules based on the LFDR would have optimal Bayes risk. That is, they would perform at least as well on average as any other decision rule with respect to any bounded loss function. Knowledge of the LFDR would require knowledge not only of the proportion of null hypotheses that are true but also the distribution of the reduced data under the alternative hypotheses. In that case, there would be no objection against relying on the LFDR derived from Bayes's theorem since frequentists by principle condition on the data in the presence of a known population of parameter values (Fisher, 1973; Wilkinson, 1977; Edwards, 1992; Kyburg and Teng, 2006; Hald, 2007, p. 36; Yuan (2009); Fraser, 2009). With that knowledge, the unquestioned applicability of the LFDR would hold regardless of the number of hypotheses that correspond to measurements. As a result, the LFDR would apply to a single comparison corresponding to a hypothesis randomly drawn from the population (Example 1) no less than to multiple comparisons spanning the entire population of hypotheses.

However, it is generally believed that the LFDR can only be adequately estimated if there are data directly related to thousands of hypotheses. For example, if data are only available for 20 genes, or, in the case study of this paper, 20 proteins, then the LFDR is not considered applicable. Indeed, empirical Bayes methods designed for several thousands of comparisons do not necessarily work as well with smaller numbers of hypotheses.

In some respects, that limitation of the empirical Bayes framework restricts the utility of multiple comparison procedures more generally. The discussions of two empirical Bayes papers spanning the last three decades (Morris, 1983a; Efron, 2010a) illustrate the consensus that very different procedures seem suitable for different numbers of comparisons. Westfall (2010) emphasized in his comment that whereas methods that control family-wise error rates (FWERs) have insufficient statistical power for very large numbers of comparisons, estimators of FDRs and LFDRs become unreliable for small numbers of comparisons. Efron (2010c) replied with a recommendation for FWER control for smaller numbers of comparisons as a substitute for empirical Bayes estimation of the FDR for larger numbers of comparisons. That conflicts with the viewpoint of Morris (1983b), another pioneer of empirical Bayes procedures, who resorted to fully Bayesian procedures for small numbers of comparisons.

The main purpose of this paper is to extend the scope of LFDR estimation to the smallest possible scale: that of a single comparison. The investigation will involve modifying a successful method of LFDR estimation and studying its relative performance in various contexts. It will be compared to fully Bayesian inference under a default prior and to the p-value interpreted inferentially with the aid of confidence distributions. The importance of the p-value in the multiple comparison framework lies in the fact that it is equal to the p-value adjusted to control an error rate when only one comparison is made. For example, with data for only a single hypothesis test, the achieved FDR, the lowest value at which the FDR has guaranteed control, is equal to the p-value (Benjamini and Hochberg, 1995).

Were such a method of small-scale LFDR estimation available for small-scale genetic association studies, the widespread publication of significant findings that could not be replicated (Morgenthaler and Thilly, 2007) might have been avoided. The reason is that LFDR estimation takes advantage of an estimate of the proportion of null hypotheses that are true, which is crucial for extremely small proportions, whereas p-values ignore that information, thereby inflating the Type I error rate of testing a hypothesis picked at random.

Example 2. For testing hundreds of thousands of genetic variants for association with disease, FWER control in the tradition of Bonferroni, Sidak (1967), and Holm (1979) often, due to the large number of tests, results in the rejection of few or no null hypotheses. The alarming number of false positives found in candidate gene studies (Morgenthaler and Thilly, 2007) at first seems to support such adjustments of p-values for the number of tests in order to control an FWER. However, the analogous history of false positives in candidate-gene studies (Ioannidis et al., 2001), in which much smaller numbers of tests were performed in each study, shows that the number of tests is not the source of the high false-positive rate. Rather, the root of the problem lies more in the small number of disease-associated variants compared to the total number of variants, irrespective of how many happen to be measured. Thus, many join the Wellcome Trust Case Control Consortium (2007) in questioning “the view that one should correct significance levels for the number of tests performed to obtain 'genome-wide significance levels."' In place of the number of tests performed, the Wellcome Trust Case Control Consortium (2007) uses the proportion of variants that are associated with disease as the prior probability of association, an approach that applies in principle even to data representing only a single variant. That proportion is thought to be between 10^{-6} and 10^{-4}, as in Example 1.

Section 2 introduces a parametric method that enables empirical Bayes inference even in the absence of multiple comparisons. Next, Section 3 derives rival posterior distributions from confidence intervals under fixed-parameter models. An application to proteomics data illustrates the empirical Bayes and confidence methods in Section 4. Section 5 compares the performance of the empirical Bayes and confidence methods for inference about a single scalar parameter value that belongs to some population of parameter values. The paper closes in Section 6 with a discussion of the resulting implications on whether empirical Bayes or confidence strategies would be more suitable in a given context.

2 Empirical Bayes methods

While methods of estimating the LFDR on the basis of nonparametric density estimators clearly cannot apply to single-comparison data (Efron, 2010b), it will be seen that fully parametric methods of LFDR estimation by maximum likelihood can do so under sufficiently simple models. Since the empirical Bayes models that define the LFDR have random parameters, the likelihood is not maximized over their values but rather over the values of the hyperparameters specifying the proportion of null hypotheses that are true and the distribution of the reduced data under the alternative hypothesis. Such parameters, if known, would entail knowledge of the LFDR (§1). More generally, the maximization of likelihood over hyperparameters is called *Type II maximum likelihood* as opposed to the Type I maximum likelihood of models that lack random parameters (Good, 1966).

2.1 Hierarchical sampling model

2.1.1 Level 1 of the model

Consider a reference set of $\tilde{N}$ populations that includes the N populations sampled. Thus, N is the number of comparisons can be made on the basis of available data. For example, $\tilde{N}$ may be the number of genes in the genome, whereas N is the number of genes on the microarray that measures gene expression or is equal to 1 if the expression of only a single gene is measured. Here, a *comparison* is understood as a hypothesis test or an effect-size estimate.

Let X_i, an observable vector of dimension n, be a random variable of a distribution P_{θ_i,λ_i}, which depends on θ_i, the parameter of interest, and on λ_i, the nuisance parameter, for all $i \in \left\{1, \dots, \tilde{N}\right\}$. Similarly, model x_j, the vector of n observations, as a realization of X_j for all $j \in \{1, \dots, N\}$.

Those data are reduced as follows. A *random statistic* U_i is a function of X_i, and an *observed statistic* u_j is a function of x_j, where the same function is applied to all $i \in \left\{1, \dots, \tilde{N}\right\}$ and to all $j \in \{1, \dots, N\}$. Thus, u_i is a realization of U_i for all $i \in \{1, \dots, N\}$.

Supposing the distribution of U_i is indexed by the *reduced parameter* δ_i, a function of θ_i and λ_i, its probability mass function or density function is denoted by $f(\bullet; \delta_i)$ for each $i \in \left\{1, \dots, \tilde{N}\right\}$. It follows that the probability mass or density of u_i is $f(u_i; \delta_i)$ for all $i \in \{1, \dots, N\}$. Without loss of generality, the ith null hypothesis is that $\theta_i = 0$ or, equivalently, $\delta_i = 0$, for any $i \in \left\{1, \ldots, \tilde{N}\right\}$.

Example 3. Suppose the expression level of each of N genes is measured for a total of n^{treat} cell cultures treated with a chemical and n^{control} cell cultures not so treated. The expression level of the ith gene is the logarithm of a measure of the abundance of mRNA in the cells and is IID $\mathrm{N}\left(\theta_i^{\text{treat}}, \lambda_i^2\right)$ within the treatment group and IID $\mathrm{N}\left(\theta_i^{\text{control}}, \lambda_i^2\right)$ within the control group, λ_i being the common standard deviation. Then T_i, the equal-variance Student t test statistic, has a noncentral t distribution with noncentrality parameter $\Delta_i = \left(\theta_i^{\text{treat}} - \theta_i^{\text{control}}\right)\left(1/n^{\text{treat}} + 1/n^{\text{control}}\right)^{-1/2}/\lambda_i$ and $n - 2 = n^{\text{treat}} + n^{\text{control}} - 2$ degrees of freedom; this is abbreviated by $T_i \sim \text{Student}\left(\Delta_i, n-2\right)$. Then $U_i = |T_i|$ is very effective for inference about $\delta_i = |\Delta_i|$. By implication, U_i is highly informative about the expression *fold change* $\exp\left|\theta_i^{\text{treat}} - \theta_i^{\text{control}}\right|$, the effect size most often estimated in reports of microarray data analysis, and about whether $\theta_i^{\text{treat}} = \theta_i^{\text{control}}$ since that is necessary and sufficient for $\delta_i = 0$. If $n^{\text{treat}} + n^{\text{control}}$ is large enough, then $T_i \dot{\sim} \mathrm{N}\left(\Delta_i, 1\right)$, which entails that U_i^2 is approximately distributed as $\chi^2\left(\delta_i^2, 1\right)$, the noncentral chi-square distribution with noncentrality parameter δ_i^2 and 1 degree of freedom.

The most common model for analyzing genetic association data has the same asymptotics.

Example 4. Example 2, continued. In order to utilize genetic models such as the additive model (Lewis, 2002) and in order to account for effects of covariates, genetic association data are typically analyzed using the Wald approximation with logistic regression, yielding the statistic T_i equal to the (Type I) maximum likelihood estimate of the log odds ratio divided by the estimated standard error of that estimate for variant i of N. The statistic $U_i = |T_i|$ is highly informative about the absolute value of the log odds ratio and whether it is equal to 0, as under the null hypothesis of no association between the genotype and the trait. For a sufficiently high number of case and control subjects, $U_i^2 \dot{\sim} \chi^2\left(\delta_i^2, 1\right)$, as in Example 3.

2.1.2 Level 2 of the model

The first level of the hierarchical model describes the variability of the expression levels of each gene or other population that corresponds to a comparison (§2.1.1). To represent variability between populations or comparisons, δ_i is now modeled as the random variable equal to 0 with probability π_0, equal to some $\delta^{(1)} \neq 0$ with probability π_1, equal to some $\delta^{(2)} \notin \left\{0, \delta^{(1)}\right\}$ with probability π_2, ..., and equal to some $\delta^{(K)} \notin \left\{0, \delta^{(K)}\right\}$ with probability for a $K \in \{1, 2, \dots\}$. The alternative-hypothesis parameters constitute ψ, a matrix with $\langle \pi_1, \dots, \pi_K \rangle$ and $\left\langle \delta^{(1)}, \dots, \delta^{(K)} \right\rangle$ as its two columns.

Then the unknown hyperparameters are π_0 and ψ, and the probability mass function or density function of X_i is the finite mixture

$$\bar{f}\left(\bullet; \pi_0, \psi\right) = \pi_0 f\left(\bullet; 0\right) + \sum_{k=1}^{K} \pi_k f\left(\bullet; \delta^{(k)}\right)$$

for all $i \in \left\{1, \dots, \tilde{N}\right\}$. The random indicator ν_i will determine whether the null hypothesis is true ($\nu_i = 1$) or false ($\nu_i = 0$) for all $i \in \left\{1, \dots, \tilde{N}\right\}$. It is assumed that $\tilde{N}$ is large enough that $P\left(\nu_i = 1\right) = \pi_0$ is approximately $\sum_{i=1}^{\tilde{N}} \nu_i / \tilde{N}$, the proportion of null hypotheses that are true.

The *local false discovery rate*, $P\left(\nu_i = 1 | U_i = u_i\right)$ by definition, is

$$\text{LFDR}\left(u_i; \pi_0, \psi\right) = \frac{P\left(\nu_i = 1\right) \bar{f}\left(u_i | \nu_i = 1; \pi_0, \psi\right)}{\bar{f}\left(u_i; \pi_0, \psi\right)} = \frac{\pi_0 f\left(u_i; 0\right)}{\bar{f}\left(u_i; \pi_0, \psi\right)}$$

by Bayes's theorem. As this LFDR is unknown only because π_0 and ψ are unknown, it may be estimated by Type II maximum likelihood, as will now be seen.

2.2 Type II maximum likelihood

The hyperparameters are estimated by $\hat{\pi}_0$ and $\hat{\psi}$, the values of π_0 and ψ at which the likelihood

$$\prod_{i=1}^{N} f\left(x_i; \pi_0, \psi\right)$$

attains its maximum subject to the constraints that $\sum_{k=1}^{K} \pi_k = 1$ and $0 \le \pi_k \le 1$. Then $\text{LFDR}\left(u_i; \hat{\pi}_0, \hat{\psi}\right)$ is the maximum likelihood estimate of the LFDR. Pawitan et al. (2005), Muralidharan (2010), and Yang and Bickel (2010) employed this method of estimating the LFDR under fully parametric finite mixtures.

To prevent overfitting in the form of excessive variance in the estimates, the value of K must be smaller for smaller values of N. For that reason, Bickel (2010d) suggested $K = 1$ when $N < 1000$. That model is simpler than those of higher values of K: the only free parameters are π_0, the approximate proportion of null hypotheses that are true, and $\delta^{(1)}$, the value of the reduced parameter indexing the alternative distribution. However, it is not simple enough for a single comparison ($N = 1$), for in that case, $\hat{\pi}_0 = 0$ almost always.

More generally, whenever N is deemed too small for reliable estimation of $\hat{\pi}_0$ with π_0 only restricted to the interval $[0, 1]$, it will be further constrained to the strictly smaller interval $\left[\pi_0^-, \pi_0^+\right]$, a proper subset of $[0, 1]$ with the specified bounds π_0^- and π_0^+ such that $0 \le \pi_0^- \le \pi_0^+ \le 1$. Thus, the proposed method guarantees that $\pi_0^- \le \hat{\pi}_0 \le \pi_0^+$ even for the lowest values of N.

In the case of $N = 1$, there is overfitting in the sense that $\hat{\pi}_0 = \pi_0^-$ almost always. Likewise, for small values of N, $\hat{\psi}$ is not an optimal estimator of ψ. Thus, improvements such as those based on predictive distributions are certainly possible (e.g., Bickel, 2011). Nonetheless, the application (§4) and simulations (§5) demonstrate that even the simple method introduced here can perform substantially better than methods that take no account of the hierarchical structure of the data. It will be seen that with certain distributions of unknown parameter values, even extremely crude estimates of the hyperparameters are preferable to no estimates at all.

To prevent problems with numerically maximizing the likelihood, the reduced parameter $\delta^{(1)}$ was constrained under the alternative hypothesis to have a lower bound of 10^{-3} for Sections 4 and 5, but none of the results was sensitive to the value of that bound.

3 Confidence methods

This section confines attention to the single-level model consisting of the model of Section 2.1.1 with fixed parameters rather than the random parameters of Section 2.1.2. The concept of confidence posterior distributions will be reviewed to set the stage for the observed confidence levels to consider as viable alternatives to LFDRs.

Let $\Theta \subseteq \mathbb{R}^1$ denote the *parameter space* of each fixed parameter value θ_i in the sense that it is the smallest set in which θ_i is known to lie. Likewise, let Λ denote the parameter space of each λ_i. Whereas the nuisance parameter λ_i may be a scalar or vector, it is assumed that the interest parameter θ_i is a scalar, i.e., that $\Theta \subseteq \mathbb{R}^1$.

Consider ϑ_i, the random variable that has probability distribution $P(\bullet; u_i)$ on Θ such that

$$P(\vartheta_i \leq \theta_i; u_i) = P_{\theta_i,\lambda_i}(U_i \geq u_i) \tag{2}$$

for all $\theta_i \in \Theta$ and $\lambda_i \in \Lambda$, where U_i is a scalar statistic determined by P_{θ_i,λ_i}, the sampling distribution of X_i introduced in Section 2.1.1. The random elements of the equation are ϑ_i on the left-hand side but U_i on the right-hand side.

The probability measure $P(\bullet; u_i)$ is the *confidence* posterior *distribution* of θ_i. The word *confidence* emphasizes the property that the interval bounded by the β_1-quantile and the β_2-quantile of ϑ_i is a $(\beta_2 - \beta_1)\, 100\%$ confidence interval in the sense that it has a $(\beta_2 - \beta_1)\, 100\%$ frequentist probability of including θ_i (Efron, 1993; Schweder and Hjort, 2002; Singh et al., 2005). While the term *posterior* correctly indicates the dependence of the parameter distribution $P(\bullet; u_i)$ on the observed statistic u_i (Bickel, 2010b,a), it is not necessarily a Bayesian posterior, a conditional prior distribution given $U_i = u_i$. For example, $P(\theta^- \leq \vartheta_i \leq \theta^+; u_i)$ is the confidence posterior probability of the hypothesis that the parameter of interest lies between the fixed values θ^- and θ^+ and yet need not correspond to any Bayesian posterior probability of the hypothesis. Polansky (2007) calls $P(\theta^- \leq \vartheta_i \leq \theta^+; u_i)$ the *observed confidence level* of the hypothesis; cf. Efron and Tibshirani (1998).

Example 5. Example 3, continued. For simplicity, the statistic is changed to $U_i = T_i$, which is useful for inference about the value of $\theta_i = \theta_i^{\text{treat}} - \theta_i^{\text{control}}$. Since $T_i \sim \text{Student}(\Delta_i, n-2)$, equation (2) implies that $\vartheta_i / \hat{\sigma}_i \sim \text{Student}(t_i, n-2)$, where $\hat{\sigma}_i$ is the typical pooled estimate of the standard error of the sample mean difference between treatment and control (Schweder and Hjort, 2002). Thus, the confidence posterior distribution of the parameter of interest is equivalent to the Bayesian posterior distribution resulting from the improper priors according to which the mean and the logarithm of the standard deviation are uniform on the real line. Coherence in the Bayesian sense would then require that the same posterior distribution be used for inference about $|\theta_i|$, e.g.,

$$\begin{aligned} P\left(|\vartheta_i| = 0; t_i\right) = P\left(\vartheta_i = 0; t_i\right) &= P\left(\vartheta_i \le 0; t_i\right) - \lim_{\epsilon \to 0+} P\left(\vartheta_i \le 0 - \epsilon; t_i\right) \\ &= P_{0,\lambda_i}\left(U_i \ge u_i\right) - \lim_{\epsilon \to 0+} P_{0+\epsilon,\lambda_i}\left(U_i \ge u_i\right) = 0. \qquad (3) \end{aligned}$$

For $\lambda_i = 1$ and large n, $\vartheta_i^2 \dot{\sim} \chi^2\left(\delta_i^2, 1\right)$, which Stein (1959) presented as the fiducial distribution for inference about θ_i^2, contrasting its interval estimates with confidence intervals.

The next example extracts a different confidence posterior distribution from the same statistical model.

Example 6. Example 3, continued. Let $U_i = |T_i|$ to draw inferences about $\theta_i = \left|\theta_i^{\text{treat}} - \theta_i^{\text{control}}\right|$. By equation (2), $P\left(\bullet; u_i\right)$, the confidence posterior distribution of ϑ_i, is defined by

$$P\left(\vartheta_i \le \theta_i; u_i\right) = P_{\theta_i,\lambda_i}\left(|T_i| \ge u_i\right).$$

Because $T_i \sim \text{Student}\left(0, n-2\right)$ under the null hypothesis that $\theta_i = 0$, the confidence posterior probability that the null hypothesis is true is equal to the usual two-sided p-value:

$$P\left(\vartheta_i = 0; u_i\right) = P\left(\vartheta_i \le 0; u_i\right) = P_{0,\lambda_i}\left(|T_i| \ge u_i\right). \qquad (4)$$

This is a clear counterexample to the observation of Polansky (2007) and Bickel (2010b) that many confidence posteriors e.g., that of Example 5, put no probability mass on any simple hypothesis.

Like the Bayesian posterior, the confidence posterior can be used to make coherent decisions given a loss function (Bickel, 2010b,a). In the metaphor of an intelligent agent, whereas the Bayesian posterior describes the decisions made by an agent committed to a particular prior distribution, the confidence posterior describes the decisions made by an agent that interprets confidence levels from a particular procedure as levels of certainty (Bickel, 2009). Thus, the confidence posterior enables direct performance comparisons between frequentist procedures and Bayesian and empirical Bayes posteriors, as will be seen in Sections 4 and 5.1.

4 Application to proteomics data

Alex Miron's lab at the Dana-Farber Cancer Institute recorded the abundance level of each of 20 plasma proteins for every woman of two breast-cancer groups (55 HER2-positive women and 35 mostly-ER/PR-positive women) and of a control group (64 healthy women) (Li, 2009). After adding the 25th percentile of the abundance levels within the control group to all abundance levels in order to ensure that the adjusted levels were positive (Bickel, 2010d), the logarithms of the adjusted levels of a given gene were modeled as quantities drawn from a normal distribution with the same variance.

In comparing each breast-cancer group to the control group, the data for each protein were reduced to the absolute value of the equal-variance t-statistic, which has a Student t distribution under the null hypothesis of no difference between groups and a noncentral Student t distribution with noncentrality parameter δ under the alternative hypothesis of a nonzero mean difference, as in Example 3.

In order to analyze the data of all proteins simultaneously, it was assumed that the reduced data of all proteins with differential abundance levels are absolute values of variates drawn from the same noncentral t distribution, the noncentrality parameter of which is denoted by δ. The assumption enabled computing $\hat{\pi}_0$ and $\hat{\delta}$, the maximum likelihood estimates of π_0 and δ, using the empirical Bayes method of Section 2.2 with the constraint that $0\% \leq \pi_0 \leq 100\%$. For comparison, the data of each protein were then analyzed individually by using the confidence and empirical Bayes methods as if it were the only protein with measured expression.

The results are summarized in Figures 1 and 2. Within each figure, the posterior probability estimates of the top-left plot are the LFDRs estimated by substituting $\hat{\pi}_0$ and $\hat{\delta}$ for π_0 and δ, with the vertical line specifying the value of $\hat{\pi}_0$. Each posterior probability of each top-right plot is the observed confidence level of the null hypothesis of equivalent abundance between cancer and control groups as recorded by equation (4). The bottom two plots of each figure report the LFDRs estimated separately for each protein by maximizing the likelihood with the constraints that $\pi_0 \geq 50\%$ (bottom-left plot) and $\pi_0 \geq 90\%$ (bottom-left plot), with the vertical lines drawn at 50% and 90%, respectively.

Since only the top-left plot of each figure represents the simultaneous use of the data for all proteins, it serves as the reference for evaluating the three methods of analyzing the data of each protein in isolation from the other data. As seen in Figure 1, the observed confidence levels closely match the simultaneously estimated LFDRs for the HER2-control group. By contrast, the individual-protein LFDR estimates come much closer than the observed confidence levels to the simultaneously estimated LFDRs for the ER/PR-control group (Figure 2). The explanation for that difference between comparisons is that the estimated proportion of equivalent-abundance proteins is low for the first group ($\hat{\pi}_0 \doteq 22\%$) but high for the second group ($\hat{\pi}_0 \doteq 89\%$).

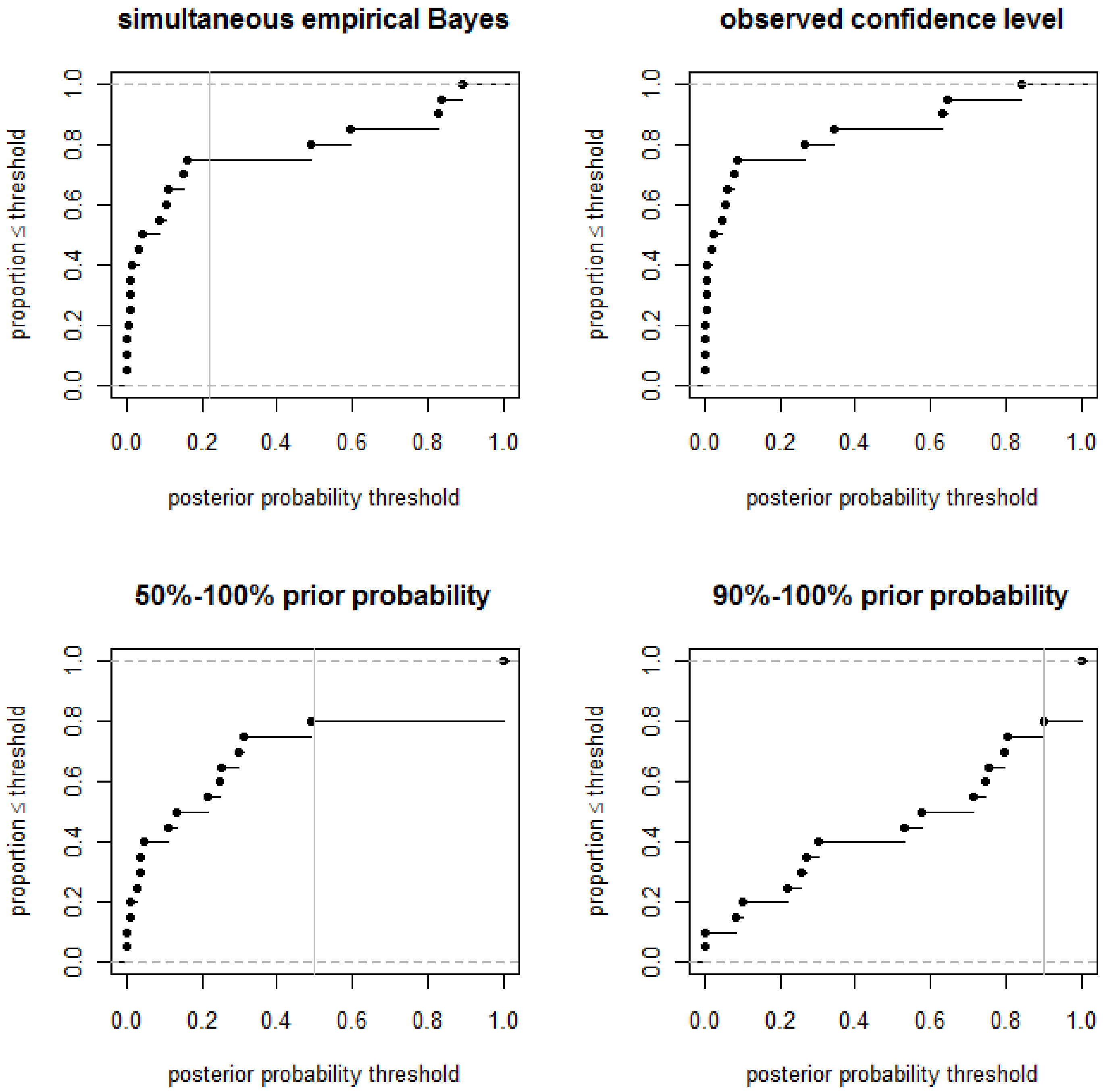

Figure 1: Empirical distribution functions of the posterior probability that a given protein has equivalent abundance between the HER2-positive and control groups. The four methods compared are described in the text.

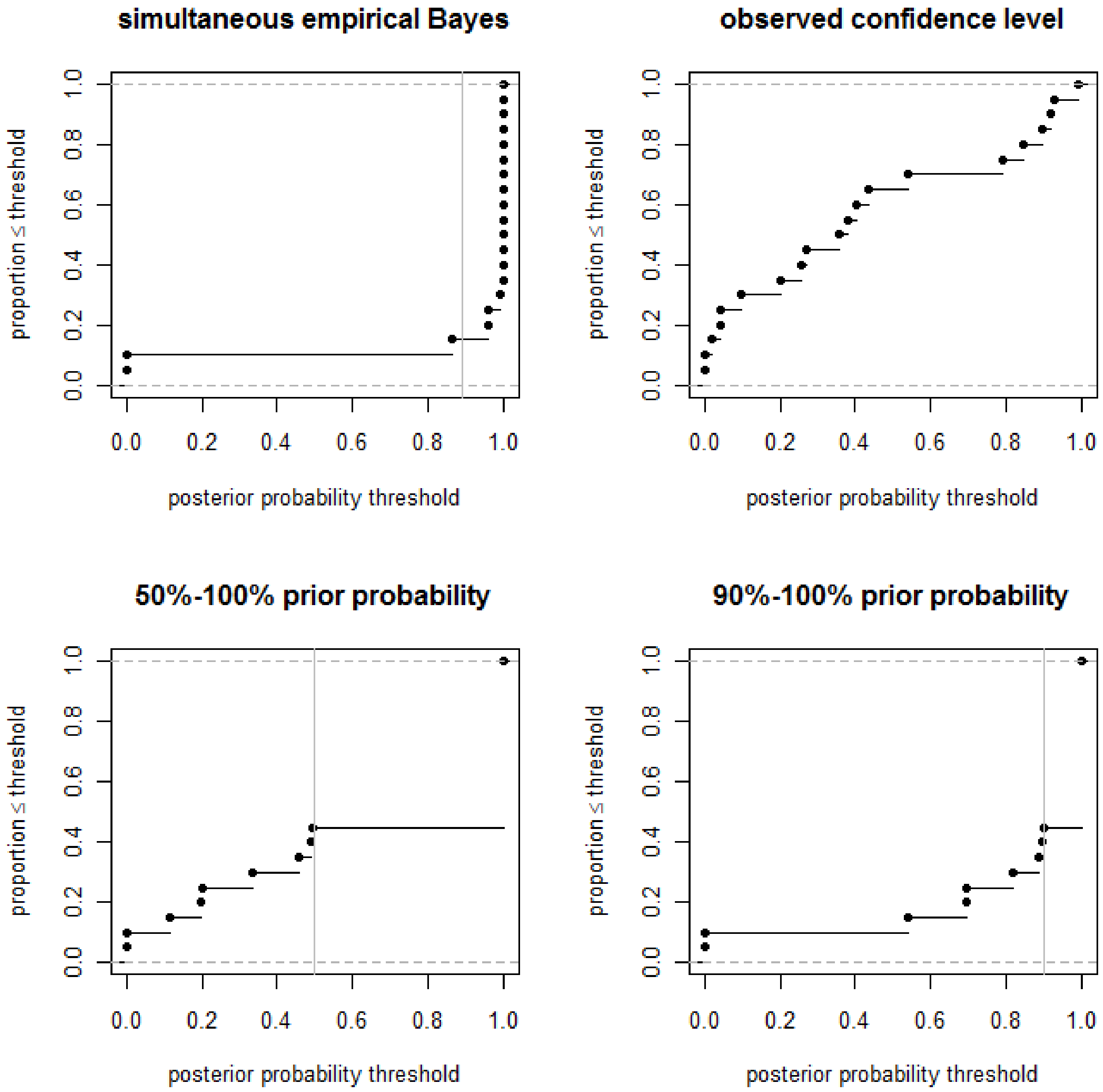

Figure 2: Empirical distribution functions of the posterior probability that a given protein has equivalent abundance between the ER/PR-positive and control groups. Each plot corresponds to a method described in the text.

5 Simulation studies

The simulation studies of the following subsections were carried out in the scenario of $T_i \sim \mathrm{N}(\Delta_i, 1)$, $U_i \sim |T_i|$, and $\delta_i = |\Delta_i|$ since it represents the asymptotics of a wide variety of situations encountered in practice, including those of protein abundance (§4), gene expression (Example 3), and genetic association (Example 4). Specifically, the test statistics were the absolute values of the realizations drawn from the normal distribution with mean $\delta = 0$ and variance 1 under the null hypothesis and from the normal distribution with mean $\delta \in \{2, 4\}$ and variance 1 under the alternative hypothesis. The mean error in estimating the truth of the null hypothesis (§5.1) or the rate at which interval estimates cover δ (§5.2) then approximated the expected error or coverage probability of each single-comparison method under the null and alternative hypotheses.

Such approximations enabled approximating the expected error and coverage probability for any proportion π_1 of null hypotheses that are false as the weighted average of the expected error or coverage probability with weight $1 - \pi_1$ for the null hypothesis and π_1 for the alternative hypothesis. This quantifies the average performance of applying each single-comparison method to data drawn from a randomly selected hypothesis.

5.1 Hypothesis testing

The posterior probability that a method attributes to the null hypothesis is its estimate of the value of the indicator ν_i that equals 1 if the null hypothesis is true or 0 if not (§2.1.2). Each method's estimation performance is here defined in terms of the mean squared error (expected quadratic loss) for two reasons. First, admissibility under quadratic loss is necessary and sufficient for certain desirable properties relevant to conditional inference (Robinson, 1979).

Second, quadratic loss is the only proper scoring rule for probabilities that (a) depends only on the difference between the estimator and estimand and (b) remains unchanged if the estimator and estimand trade places (Savage, 1971). The square root of the expected quadratic loss is easily interpreted as an average estimation error.

The present adoption of the confidence posterior probability of equation (4) is equivalent to interpreting the p-value as an estimate of the indicator of whether the null hypothesis is true. The p-value used this way does not require a significance threshold and can dominate estimates defined to equal 0 if the p-value is below such a threshold and equal to 1 otherwise (Hwang et al., 1992). Fixed-probability tails will be more appropriate for constructing the confidence intervals of Section 5.2 since it, unlike the present section, in effect imposes a 0-1 loss function (Robinson, 1979).

On the basis of 100 realizations of the statistic drawn from each of the three normal distributions $\mathrm{N}(0,1)$, $\mathrm{N}(2,1)$, and $\mathrm{N}(4,1)$, Figures 3 and 4 compare the mean quadratic loss of several methods of hypothesis testing in the general form of assigning posterior probability to the null hypothesis. The vertical lines are drawn at $\pi_1 = 50\%$. The *0% posterior probability* represents any method that necessarily assigns no probability mass to the simple null hypothesis, including improper-prior Bayesian updating and all other methods yielding posterior density functions (Example 5). The *observed confidence level* is the confidence posterior probability given by equation (4) with infinite degrees of freedom. Each of the four methods of estimating the LFDR imposes a different constraint on π_0 when maximizing the likelihood.

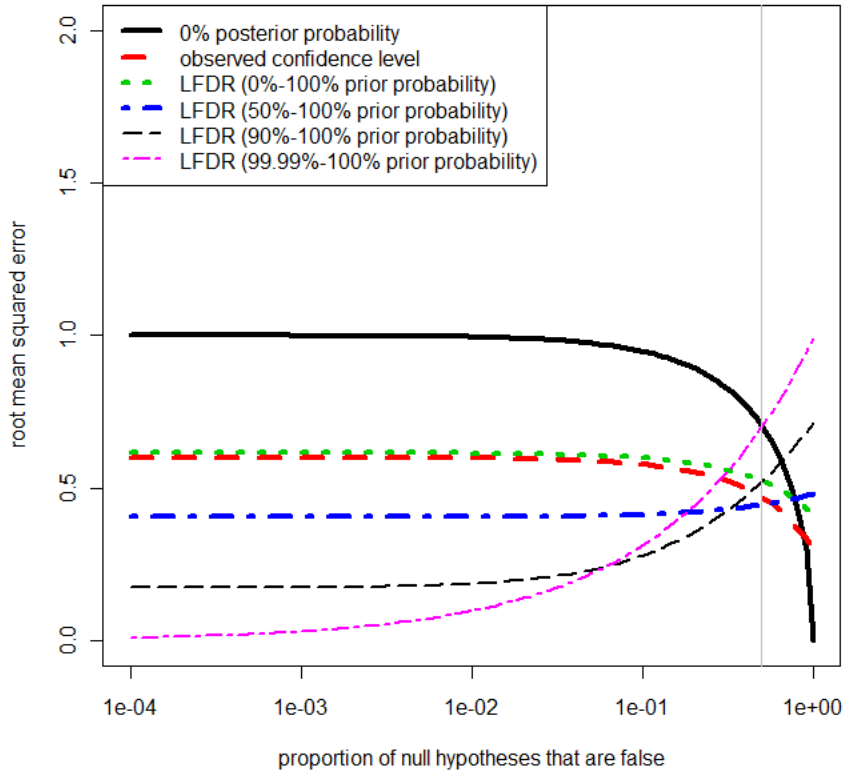

Figure 3: Square root of the mean quadratic error of the (estimated) posterior probabilities of null hypothesis truth versus $\pi_1 = 1 - \pi_0$. Reduced data were simulated from the unit-variance normal distributions of means 0 (true null hypothesis) and 2 (false null hypothesis).

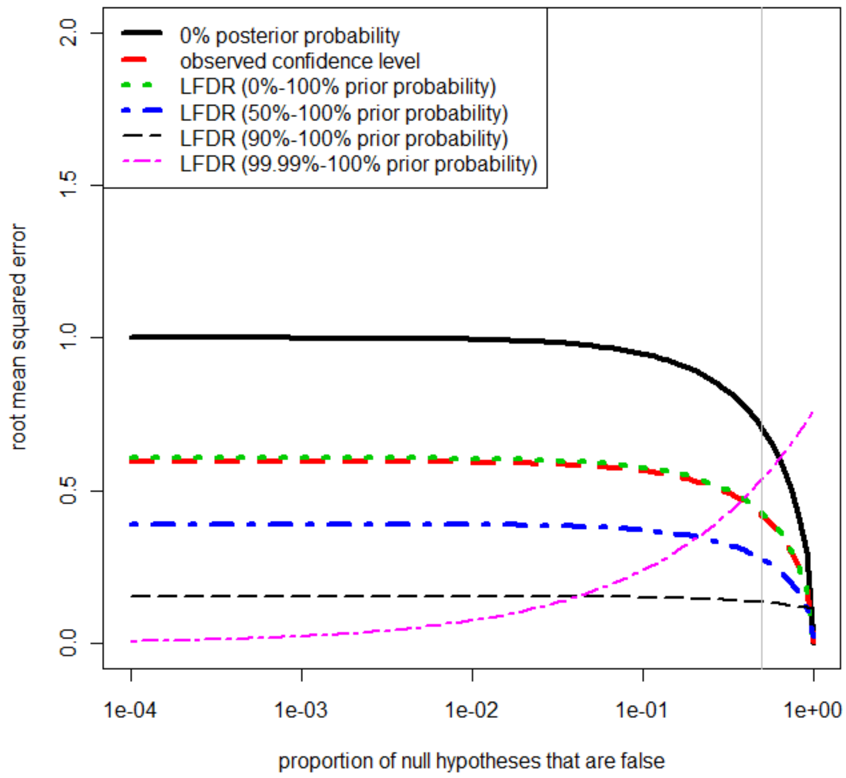

Figure 4: Square root of the mean quadratic error of the (estimated) posterior probabilities of null hypothesis truth versus $\pi_1 = 1-\pi_0$. Reduced data were simulated from the unit-variance normal distributions of means 0 (true null hypothesis) and 4 (false null hypothesis).

5.2 Effect-size estimation

An interval estimate of the effect size $|\delta|$ is the interval between two quantiles of a posterior distribution of $|\delta|$, whether a confidence posterior, a Bayesian posterior, or an empirical Bayes posterior. For example, the central or equal-tail $(1-\alpha)\,100\%$ confidence interval corresponding to a confidence posterior is the interval between its $\alpha/2$ and $1-\alpha/2$ quantiles. The coverage rate of an interval estimate is its probability of including the true value of the interest parameter, $|\delta|$ in the case of the simulation studies.

Figure 5 displays the coverage rates of the equal-tail 95% interval estimates for simulating 800 observed test statistics from the null distribution and another 800 from the alternative distribution with $\delta = 2$. The displayed coverage rates are visually indistinguishable from those instead using 800 draws from the $\delta = 4$ distribution.

The six posterior distributions of Figure 5 are those of Section 5.1, again with the vertical line at $\pi_1 = 50\%$. The improper Bayesian posterior induced by the uniform prior distribution of δ represents the class of 0%-posterior methods (Example 5). Its interval estimates were criticized by Stein (1959) and Wilkinson (1977) in favor of the confidence intervals of Figure 5. Its assignment of 0% posterior probability to the null hypothesis is evident from equation (3).

6 Discussion and conclusions

The proposed method of constraining π_0 requires no more subjective input than the popular methods of estimating the LFDR that rely on nonparametric density estimation: they depend on the assumption that π_0 be greater than about 90% (Efron, 2004). With sufficiently high choices of π_0, all such methods tend to be conservative.

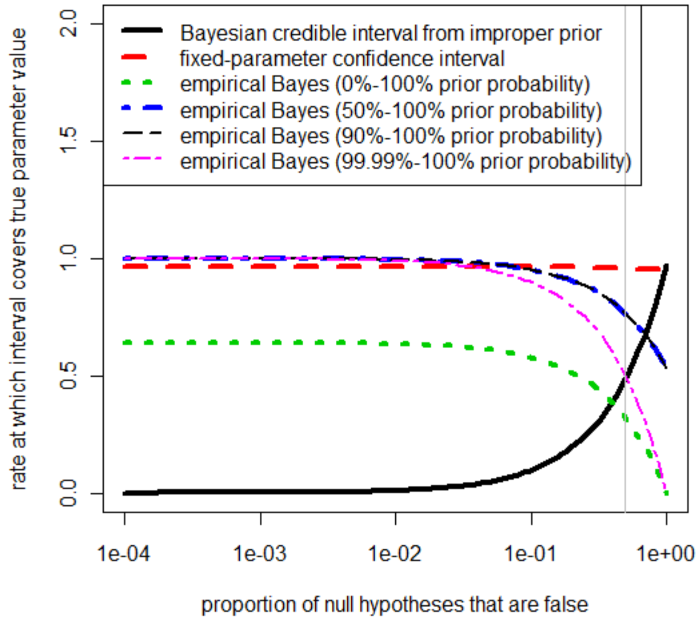

Figure 5: Proportion of 95% interval estimates that include the true value of the mean versus $\pi_1 = 1 - \pi_0$. Reduced data were simulated from the unit-variance normal distributions of means 0 (true null hypothesis) and 2 (false null hypothesis). The 50-100% and 90%-100% curves coincide.

The objection may be raised that all such choices are unnecessary given the guaranteed coverage rates of fixed-parameter confidence intervals. Indeed, although Bayesian and empirical Bayes methods can cover the true parameter at slightly higher rates, they can also have much worse coverage than confidence intervals. For example, empirical Bayes intervals based on LFDR estimation have poor coverage at high values of π_1 (Figure 5).

However, the main advantage of LFDR-based interval estimates over fixed-parameter confidence intervals lies not in the potential increase in the coverage rate but rather in the striking reduction in their width (Ghosh, 2009; Efron, 2010b; Bickel, 2010c). That is especially true for lower values of π_1, as can be seen from the greater and greater concentration of posterior probability mass at the null hypothesis as $\pi_1 \to 0$ (Figures 3 and 4). Whenever the posterior probability of the null hypothesis is at least 97.5%, which happens with close to 100% frequency for high values of the lower bound π_0^-, the 95% interval estimate is $[0, 0]$. That interval has zero width and yet will cover the true value at a rate of $1 - \pi_1$, the proportion of null hypotheses ($\theta_i = 0$) that are true.

The value of π_1 also determines whether the LFDR approach performs better or worse than the confidence approach in the context of inferring whether or not a null hypothesis is true. For $\pi_1 \dot{\leq} 10\%$, there is substantial improvement in inference even when π_0^- is far from $1 - \pi_1$ (Figures 2, 3, and 4).

Among others, Lindley (1957) and Berger and Sellke (1987) contrasted Bayesian posterior probabilities of simple null hypotheses with p-values before the LFDR was defined. The results of Berger and Sellke (1987) hold without their reliance on the misinterpretation of the p-value as a Bayesian posterior probability since, in confidence-posterior decision theory (Bickel, 2010b,a), the two-sided p-value can be a legitimate confidence posterior probability (4). Berger and Sellke (1987) found that the p-value can be far from the actual error rate, which necessarily depends on π_1, the proportion of null hypotheses that are false, whether or not that proportion is known. That, however, is insufficient for concluding that Bayesian testing is superior: in low-information situations, Bayesian posterior probabilities will also be far from those that would be computed with knowledge of π_1 and other model parameters. For the practical scientist who does not want to know about error rates but instead whether or not the null hypothesis is true, the more important criterion is whether Bayesian posterior probabilities or p-values come closer to ν_i, the indicator of the truth of the ith null hypothesis.

Using that criterion actually favored the p-value as an observed confidence level over the empirical Bayes methods for $\pi_1 \dot{\geq} 50\%$ (Figures 1, 3, and 4). That largely vindicates the use of confidence-based methods when all that is known about the parameter of interest is encoded either in the model or in the test appropriate for a plausible null hypothesis (§3).

Nonetheless, even with the vague information that the hypothesis tested belongs to a relevant class in which most null hypotheses are true, rough guesses at π_0^- can bring notable improvements in inference accuracy. An extreme case is that of genetic association studies (Example 2), for which $\pi_1^- = 10^{-6}$ and $\pi_1^+ = 10^{-4}$ are reasonable lower and upper bounds of the proportion of SNPs associated with a given disease (Wellcome Trust Case Control Consortium, 2007).

The need to consider π_1 when making statistical inferences cannot be avoided by running algorithms that automatically control the FDR or FWER. The fundamental difference between the LFDR and the FDR is exposed at lower numbers of comparisons and especially at the single-comparison scale. Since FDR control reduces to standard hypothesis testing when there is only a single test (Benjamini and Hochberg, 1995), the achieved FDR, like any achieved FWER, is the unadjusted p-value and thus is suitable in the same high-π_1 situations.

Acknowledgments

The Biobase (Gentleman et al., 2004) package of R (R Development Core Team, 2008) facilitated the computations.This research was partially supported by the Canada Foundation for Innovation, by the Ministry of Research and Innovation of Ontario, and by the Faculty of Medicine of the University of Ottawa.

References

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57, 289–300.

Berger, J. O., Sellke, T., 1987. Testing a point null hypothesis: The irreconcilability of p values and evidence. Journal of the American Statistical Association 82 (397), 112–122.

Bickel, D. R., 2009. A frequentist framework of inductive reasoning. Technical Report, Ottawa Institute of Systems Biology, arXiv:math.ST/0602377.

Bickel, D. R., 2010a. Coherent frequentism: A decision theory based on confidence sets. To appear in Communications in Statistics - Theory and Methods (accepted 22 November 2010); preprint available from arXiv:0907.0139.

Bickel, D. R., 2010b. Estimating the null distribution to adjust observed confidence levels for genome-scale screening. Biometrics, DOI: 10.1111/j.1541-0420.2010.01491.x.

Bickel, D. R., 2010c. Large-scale interval and point estimates from an empirical Bayes extension of confidence posteriors. Technical Report, Ottawa Institute of Systems Biology, arXiv:1012.6033.

Bickel, D. R., 2010d. Minimum description length methods of medium-scale simultaneous inference. Technical Report, Ottawa Institute of Systems Biology, arXiv:1009.5981.

Bickel, D. R., 2011. A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. To appear in the Canadian Journal of Statistics (accepted 21 April 2011); preprint available from arXiv:1010.0694.

Edwards, A. W. F., 1992. Likelihood. Johns Hopkins Press, Baltimore.

Efron, B., 1993. Bayes and likelihood calculations from confidence intervals. Biometrika 80, 3–26.

Efron, B., 2004. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association 99 (465), 96–104.

Efron, B., 2010a. Correlated z-values and the accuracy of large-scale statistical estimates. Journal of the American Statistical Association 105 (491), 1042–1055.

Efron, B., 2010b. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge University Press.

Efron, B., 2010c. Rejoinder to comments on B. Efron, "Correlated z-values and the accuracy of large-scale statistical estimates". Journal of the American Statistical Association 105 (491), 1067–1069.

Efron, B., Tibshirani, R., 1998. The problem of regions. Annals of Statistics 26 (5), 1687–1718.

Efron, B., Tibshirani, R., Storey, J. D., Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. J. Am. Stat. Assoc. 96 (456), 1151–1160.

Fisher, R. A., 1973. Statistical Methods and Scientific Inference. Hafner Press, New York.

Fraser, D. A. S., 2009. Is Bayes posterior just quick and dirty confidence? Technical Report, Department of Statistics, University of Toronto.

Gentleman, R. C., Carey, V. J., Bates, D. M., et al., 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 5, R80.

Ghosh, D., 2009. Empirical Bayes methods for estimation and confidence intervals in high-dimensional problems. Statistica Sinica 19 (1), 125–143.

Good, I. J., 1966. How to Estimate Probabilities. IMA Journal of Applied Mathematics 2 (4), 364–383.

Hald, A., 2007. A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935. Springer, New York.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6 (2), 65–70.

Hwang, J. T., Casella, G., Robert, C., Wells, M. T., Farrell, R. H., 1992. Estimation of accuracy in testing. The Annals of Statistics 20 (1), 490–509.

Ioannidis, J., Ntzani, E., Trikalinos, T., Contopoulos-Ioannidis, D., 2001. Replication validity of genetic association studies. Nature Genetics 29 (3), 306–309.

Kyburg, H. E., Teng, C. M., 2006. Nonmonotonic logic and statistical inference. Computational Intelligence 22 (1), 26–51.

Lewis, C. M., 2002. Genetic association studies: Design, analysis and interpretation. Briefings in Bioinformatics 3 (2), 146 –153.

Li, X., 2009. ProData. Bioconductor.org documentation for the ProData package.

Lindley, D., 1957. A statistical paradox. Biometrika 44 (1-2), 187–192.

Lindley, D. V., 1957. A statistical paradox. Biometrika 44 (1/2), pp. 187–192.

Morgenthaler, S., Thilly, W. G., 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutation Research 615 (1-2), 28–56.

Morris, C. N., 1983a. Parametric empirical Bayes inference: Theory and applications. Journal of the American Statistical Association 78 (381), 47–55.

Morris, C. N., 1983b. Parametric empirical Bayes inference: Theory and applications: Rejoinder. Journal of the American Statistical Association 78 (381), 63–65.

Muralidharan, O., 2010. An empirical Bayes mixture method for effect size and false discovery rate estimation. Annals of Applied Statistics 4, 422–438.

Pawitan, Y., Murthy, K., Michiels, S., Ploner, A., 2005. Erratum: Bias in the estimation of false discovery rate in microarray studies (bioinformatics) vol. 21(20) (3865-3872)). Bioinformatics 21 (24), 4435.

Polansky, A. M., 2007. Observed Confidence Levels: Theory and Application. Chapman and Hall, New York.

R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Robinson, G. K., 1979. Conditional properties of statistical procedures. The Annals of Statistics 7 (4), 742–755.

Savage, L. J., 1971. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association 66 (336), pp. 783–801.

Schweder, T., Hjort, N. L., 2002. Confidence and likelihood. Scandinavian Journal of Statistics 29 (2), 309–332.

Sidak, Z., 1967. Rectangular confidence regions for means of multivariate normal distributions. Journal of the American Statistical Association 62 (318), 626–633.

Singh, K., Xie, M., Strawderman, W. E., 2005. Combining information from independent sources through confidence distributions. Annals of Statistics 33 (1), 159–183.

Stein, C., 1959. An example of wide discrepancy between fiducial and confidence intervals. The Annals of Mathematical Statistics 30 (4), 877–880.

Wellcome Trust Case Control Consortium, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447 (7145), 661–678.

Westfall, P. H., 2010. Comment on B. Efron, "Correlated z-values and the accuracy of large-scale statistical estimates". Journal of the American Statistical Association 105 (491), 1063–1066.

Wilkinson, G. N., 1977. On resolving the controversy in statistical inference (with discussion). Journal of the Royal Statistical Society. Series B (Methodological) 39 (2), 119–171.

Yang, Y., Bickel, D. R., 2010. Minimum description length and empirical Bayes methods of identifying snps associated with disease. Technical Report, Ottawa Institute of Systems Biology, COBRA Preprint Series, Article 74, available at biostats.bepress.com/cobra/ps/art74.

Yuan, B., 2009. Bayesian frequentist hybrid inference. Annals of Statistics 37, 2458–2501.